Central European Journal of Operations Research (2025) 33:1503–1528
https://doi.org/10.1007/s10100-024-00932-1

Does cross-validation work in telling rankings apart?

Balázs R. Sziklai[1,2] · **Máté Baranyi**[3] · **Károly Héberger**[4]

Accepted: 10 August 2024 / Published online: 29 August 2024

Abstract
Although cross-validation (CV) is a standard technique in machine learning and data science, its efficacy remains largely unexplored in ranking environments. When evaluating the significance of differences, cross-validation is typically coupled with statistical testing, such as the Dietterich, Alpaydin, or Wilcoxon test. In this paper, we evaluate the power and false positive error rate of the Dietterich, Alpaydin, and Wilcoxon statistical tests combined with cross-validation each operating with folds ranging from 5 to 10, resulting in a total of 18 variants. Our testing setup utilizes a ranking framework, similar to the Sum of Ranking Differences (SRD) statistical procedure: we assume the existence of a reference ranking, and distances are measured in L_1-norm. We test the methods under artificial scenarios as well as on real data borrowed from sports and chemistry. The choice of the optimal CV test method depends on preferences related to the minimization of errors in type I and II cases, the size of the input, and anticipated patterns in the data. Among the investigated input sizes, the Wilcoxon method with eight folds proved to be the most effective, although its performance in type I situations is subpar. While the Dietterich and Alpaydin methods excel in type I situations, they perform poorly in type II scenarios. The inadequate performances of these tests raises questions about their efficacy outside of ranking environments too.

Keywords *k*-fold cross-validation · Rankings · Sum of ranking differences · Wilcoxon test · Alpaydin test · Leave-many-out · Multi-criteria decision-making

1 Introduction

Cross-validation (CV) is a statistical technique frequently employed, among others, in machine learning to assess the performance of predictive models and minimize the risk of overfitting. It is often coupled with statistical testing to measure significance and compare performances, with Dietterich (1998) and Alpaydin (1999) tests being common choices.

Máté Baranyi and Károly Héberger have contributed equally to this work.

Extended author information available on the last page of the article

In this paper, we analyze the efficacy of statistical testing combined with cross-validation in a specific setting. Our setup has three peculiarities: (i) we study the tests in a ranking environment, (ii) we assume the existence of a reference, and (iii) distances between rankings are measured using the L_1-norm. While these assumptions may seem restrictive, we argue that the conclusions drawn hold relevance for a broader audience.

Firstly, ranking objects is one of the most commonly applied computational tasks. In social sciences, universities are ranked by excellence, sports teams are ranked according to various performance measures, politicians are ranked by their popularity, etc. In machine learning, query results are ranked by search engines, features are ranked during feature selection, algorithms are ranked by their performance, etc.

Secondly, there is often a reference through which the solutions are compared to each other. In content recommendation, this can be the user for whom we would like to generate suggestions; in image search, the queried image itself serves as a reference; in machine learning, the test dataset is used as a reference for algorithms refined on the training dataset; in preference elicitation, such as in the Mallows model (1957), consumer choices are compared through a fixed reference.

Choosing the L_1-norm as the distance metric is convenient as the exact same setup is employed in the Sum of Ranking Differences (SRD) statistical test. SRD, introduced by Héberger (2010), is a relatively novel procedure that is especially suitable for comparing methods in multi-criteria optimization environments. In recent years, there have been many published papers, with topics ranging from machine learning (Moorthy et al. 2017), through multi-criteria decision-making (Gere et al. 2021) and pharmacology (Vajna et al. 2012), to political science (Sziklai and Héberger 2020), and even sports (West 2018), that apply or further extend the technique, showing its versatility.

SRD compares methods or solutions through a reference by converting the input data into rankings, and then calculating the distance in L_1-norm (Spearman's footrule if no ties are present) of each method's ranking and the reference ranking. The latter can be an external gold standard or an aggregate from the data, *cf.* ref. (Héberger 2010). The SRD algorithm contains two validation steps:

- In the so-called permutation test, the methods' scores are compared to the SRD scores of random rankings. For a detailed overview of this method, see ref. Héberger and Kollár-Hunek (2011).
- In the second step, the results are cross-validated by re-sampling the data: Uncertainties are assigned to the SRD scores, which are deterministic by nature, using cross-validation. This approach also allows for grouping and comparing the methods in a different way than the pure SRD scores.

Our goal is to explore the efficacy of statistical tests coupled with cross-validation in a ranking environment, but our investigations were inspired by the second validation step of the Sum of Ranking Differences. Throughout the paper, we use this framework to study the problem. However, the only practical implications lie in the use of a reference ranking and the distance metric.

SRD scores can be seen as a metric of the mean absolute error (on a rank scale) to the reference. It is easy to draw a parallel with machine learning, where the CV of the chosen error metric is a standard technique. There are a lot of approaches in the literature about how to make conclusions from the many error scores calculated on different subsamples of the data. Not all of these can easily be adapted to ranking frameworks.

Originally, by Kollár-Hunek and Héberger (2013), the Wilcoxon signed-rank test (henceforward Wilcoxon test) was proposed for CV purposes. Dietterich's CV t-test (Dietterich 1998) and Alpaydin's CV $\mathcal{F}$-test (Alpaydin 1999) are popular regarding cross-validated machine learning algorithms. An original contribution of this paper is the adaptation of these methods to the ranking framework and testing of their performance. To obtain a complete picture, we constructed nine scenarios, that is, typical situations with different ranking data structures that CV methods could face. In each iteration step, we generated a pair of rankings from various distributions (data-generating processes) and observed how often the methods rejected that the two rankings came from the same distribution. In scenarios where the rankings were generated from the same distribution, the type I error of the hybrid tests could be assessed; whereas when they were generated from different distributions, the type II error.

The scenario analysis reveals a mixed picture. Different CV methods prevail under different circumstances (type I/II errors, input size, data structure). The Wilcoxon method with eight folds seems to be the best compromise, but its 7-fold version is also a viable alternative. Establishing a performance benchmark for the methods, is important for two reasons. Firstly, the users can better understand the potential implications of relying on these statistical methods. Secondly, it provides a reference point for evaluating any new method proposed in the future.

All statistical tests that are presented in this paper are featured as a validation option in the statistical software package, rSRD—an implementation of SRD, downloadable from the Comprehensive R Archive Network (CRAN) (Staudacher et al. 2023)—which further adds to the significance of this study.

2 Literature overview

Solutions that employ ranking techniques are very common in statistical data analysis. Applications are ranging from marketing and advertisement research (Lin 2010) through belief revision (Hild and Spohn 2008), sport (Orbán-Mihálykó et al. 2023) to feature selection (Alaiz-Rodríguez and Parnell 2020) and matching (Jiang et al. 2021). Here we review the literature from three perspectives: distance, uncertainty, and aggregation of rankings. The distinction is somewhat arbitrary as these topics often intersect.

Distance of rankings Rankings are evaluated and compared by various measures, among which Spearman's footrule (Manhattan distance between rankings) is quite common (Jiang et al. 2021; Kumar and Vassilvitskii 2010). Note that SRD is a

generalization of Spearman's footrule, that can deal with ties, and where one of the rankings (the reference) is fixed. The Kendall's tau, Cayley and Spearman distances are also popular choices (Lin 2010; Brandenburg et al. 2013; Yu et al. 2019).

Sometimes, rankings are modeled in a parametric way built on top of one of the many available distance measures, *e.g.*, Spearman's footrule; and these models are further generalized and combined with different ones. Probably the best-known is Mallows' ranking model (1957), which has been the subject of intensive study. For example, recently Vitelli et al. (2017) introduced a new tractable method for Bayesian inference in Mallows' models. Švendová and Schimek (2017) proposes an indirect inference approach to estimate the latent signal parameters that might be causal for a set of observed rankings obtained from several assessors. Lee and Yu (2012) formulate models by considering weighted distances which allow different weights for different ranks. In their earlier paper (Lee and Yu 2010) they combine a tree model and the existing distance-based models to build a model that can handle more complexity. Xu et al. (2018) propose a general angle-based model for ranking data, of which a distance-based model with Spearman's distance can be seen as a special case. Negahban et al. (2018) focuses on the multinomial logit model for the representation learning of rankings in a purely probabilistic sense, instead of relying on distance metrics.

Weighted distances are also widely utilized. In various applications, such as web queries or sports rankings, top positions hold special importance (Kumar and Vassilvitskii 2010). In certain cases, even the last ranks can provide meaningful information, while the intermediate rankings may be relatively arbitrary (Abonyi et al. 2023). Weighted distances have already been integrated into the SRD framework (Gere et al. 2022). Determining the appropriate weights for different positions is inherently subjective and can introduce a potential for manipulating the results. Employing a fixed weight function can help mitigate this issue. For example, Sziklai et al. (2022) uses the natural logarithm to weight rank inversions.

Uncertainty of rankings There are many approaches to measure the uncertainty among many rankings, but the most direct way to do so is to calculate the standard deviation of the rankings as if they were simply real vectors. For rankings $\pi_1, \dots, \pi_m$ of n elements, it is $\mathrm{SD} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(\pi_i - \overline{\pi}\right)^2}$, where $\overline{\pi}$ is simply the mean of the sample (or objects) rankings as vectors. This L_2 approach has been applied in many papers, *e.g.* in Refs. Falivene et al. (2010), Palmer et al. (2009), Triantafilis et al. (2001), Farshadfar and Amiri (2016). Aside from the above metric, Barlow and Ballin (1976) measure uncertainty with another metric taken from information theory. Rosander (1936) calculates the standard error of the mean ranking based on the rank-order correlation. In Refs. Lockwood et al. (2002), Zampetakis and Moustakis (2010), the uncertainty of rankings is considered using Bayesian modeling of the raw data, from where the rankings originate. Zuk et al. (2007) tackle the uncertainty in a ranking by introducing noise to the raw data, and comparing the original ranking to the noisy one with Top-k-list overlap and Kendall's Tau measure.

Ranking aggregation Aggregating rankings to obtain an optimal one is a problem naturally arising in many fields of science, from multi-criteria decision-making through marketing to social choice. Depending on the field and desired objective different methods were developed, from intuitive heuristics like the Borda rule to stochastic optimization algorithms and probabilistic methods. The closest to our purpose the probabilistic methods are, like the Mallows (1957), Plackett-Luce (1975) and Thurstone (Lin 2010) models which quantify the relation of each object-pair in terms of probabilities. The drawback of these methods is that they pose assumptions over the underlying distributions. For instance, Thurstone assumes that each object-pair follows a bivariate normal distribution. Since the unknown parameters are unidentifiable, further assumptions (like setting the standard deviation to 1) are needed to compute the solution. Other aggregation methods also suffer from computational problems. For instance, the Kendall tau distance is known to be computationally intractable for total orders (Bartholdi et al. 1989). Nevertheless, many recent papers utilize these models, and they appear to be both flexible, good to fit the data, and computationally tractable. See for instance the works of Mollica and Tardella (2014), Mollica and Tardella (2017), Qian and Yu (2019), Irurozki et al. (2016), Crispino et al. (2023), Gyarmati et al. (2023).

The Bayesian framework presented in Lockwood et al. (2002) addresses the challenge of determining an optimal ranking. Meanwhile, Tavanaei et al. (2018) consider rank uncertainty through parameterized weightings in their rank aggregation process. Tehrani et al. (2012) employ the Choquet integral as an underlying model for representing ranking before the aggregation. Lastly, Volkovs and Zemel (2014) introduce probabilistic models, leveraging a multinomial generative process, for preference aggregation in both unsupervised and supervised settings.

The aforementioned references usually deal with the uncertainty of rank-pairs, or see the uncertainty in the raw data itself. A distinctive contribution of this paper is that we grasp the uncertainty of a single ranking by looking at its partial rankings coming from manyfold CV, which in turn enables us to statistically test whether two rankings come from the same distribution.

3 Methodology

SRD requires the input data to be arranged in a matrix form: rows $(1, \ldots, n)$ represent objects (measured items, *e.g.* features, properties, or hidden characteristics, etc., depending on the use case), whereas columns $(1, \ldots, m)$ represent the variables (measurement methods, models or solutions, etc.) to be compared. In general, the input is a real-valued matrix $A \in \mathbb{R}^{n \times m}$, although any input can be considered which can be transformed into ordinal scale. There is one designated column containing a reference value for each object. These can be a previously established gold standard, an estimation, or even an aggregation from the input data (this last technique is also called *data fusion*). The input matrix is transformed into a ranking matrix by ranking the values in each column from the smallest to the largest element. The resulting $n \times (m + 1)$ matrix contains m model rankings $\pi_1, \ldots, \pi_m$ associated to the variables and a reference ranking π_r. From now on, we will use the term "model ranking" for

the rank-transformed columns instead of variable (ranking) because the latter term has a more ambiguous meaning.

3.1 The metric SRD is built upon

The distance metric applied in the SRD framework is simply the L_1 norm or city block (Manhattan) distance of the rankings; it is Spearman's footrule if no ties are present in the ranking. For rankings π_1, π_2, we will denote it with $d(\pi_1, \pi_2) := \sum_{\ell=1}^{n} |\pi_1(\ell) - \pi_2(\ell)|$. In particular, SRD is the distance of π_1 from the reference ranking,

$$\mathrm{SRD}(\pi_1) := d(\pi_1, \pi_r).$$

In other words, SRD compares the rankings of two different models (π_1, π_2), through the difference to a common reference ranking (π_r).

Properties of Spearman's footrule have been intensively studied over the symmetric group S_n, which contains the permutations over $1, \dots, n$, in other words, rankings without ties. The maximal distance is easy to compute:

$$M := \max_{i,j} d(\pi_i, \pi_j) = \begin{cases} \frac{n^2}{2} & \text{if n is even} \\ \frac{n^2-1}{2} & \text{if n is odd} \end{cases} \tag{1}$$

The distance is right-invariant to the composition operation, meaning

$$d(\pi_i, \pi_j) = d(\pi_i\sigma, \pi_j\sigma) \quad \forall \sigma \in S_n.$$

For convenience and interpretability, we usually normalize the distance by the maximal distance of Eq. 1:

$$\underline{SRD}(\pi) := \mathrm{SRD}(\pi)/M.$$

The normalized $\underline{SRD}$ values also make comparison possible if the numbers of objects are different.

Diaconis and Graham (1977) showed that the distance is asymptotically normally distributed (as $n \to \infty$) if we choose two permutations uniformly from S_n:

$$d(\cdot,\cdot) \sim \mathcal{N}\left(\frac{1}{3}n^2 + \mathcal{O}(n),\ \sqrt{\frac{2}{45}n^3 + \mathcal{O}(n^2)}\right).$$

Due to right-invariance, the distribution is the same if we fix one of the permutations, as it happens in the SRD framework where we have a fixed reference ranking. After normalization:

$$\underline{SRD}(\cdot) \sim \mathcal{N}\left(\frac{2}{3},\ \sqrt{\frac{8}{45n}}\right). \tag{2}$$

Based on the above properties, a hypothesis test can be created to answer the question: Is a specific model ranking (π_{model}) close enough to the reference ranking (π_r)? The null hypothesis H_0 is that the ranking is selected randomly (with uniform distribution) from S_n. The test statistic is $\underline{SRD}(\pi_{model})$, which is asymptotically normally distributed for large n under H_0 by Eq. 2. If ties are present and n is small, the exact discrete distribution should be used (Kollár-Hunek and Héberger 2013), while for large n, the empirical distribution can be simulated (Staudacher et al. 2023).

3.2 Uncertainty of SRD

As already mentioned, SRD values are deterministic by nature, but with CV, we can assign uncertainty to them in order to compare two rankings with each other. During CV we sample the original input to create folds where the SRD computation is repeated.

3.2.1 Wilcoxon signed-rank test

The Wilcoxon signed-rank test (1945) was proposed for CV purposes in (Kollár-Hunek and Héberger 2013). Take random subsets $(A_1, \ldots, A_k)$ from the rows of size $n - \lceil n/k \rceil$, where n denotes the rows/objects, while k stands for the number of folds. In other words, we randomly leave out $\lceil n/k \rceil$ number of rows in each fold. Let $\underline{SRD}_{j,i}$ denote the $\underline{SRD}$ (from the reference) on fold A_i for the (model) ranking π_j. This results in a paired sample of size k for the two models (π_1, π_2) to compare:

$$\left\{ \left(\underline{SRD}_{1,1}, \underline{SRD}_{2,1} \right), \left(\underline{SRD}_{1,2}, \underline{SRD}_{2,2} \right), \ldots, \left(\underline{SRD}_{1,k}, \underline{SRD}_{2,k} \right) \right\}.$$

To this, we apply the signed-rank test. We take the sub-sample (of size $k_r < k$) containing the non-zero absolute differences $|\underline{SRD}_{1,i} - \underline{SRD}_{2,i}|$, and then rank this subsample. Let R_i be the rank of the ith fold in this subsample. Then, the test-statistic W comes from

$$W^+ = \sum_{i=1}^{k} \mathbb{I}\left(\underline{SRD}_{1,i} > \underline{SRD}_{2,i} \right) \cdot R_i, \quad \text{and} \quad W^- = \sum_{i=1}^{k} \mathbb{I}\left(\underline{SRD}_{1,i} < \underline{SRD}_{2,i} \right) \cdot R_i,$$

where $\mathbb{I}()$ is the indicator function, which returns with value 1 if its argument evaluates true, otherwise it returns with zero. W can be $\min(W^+, W^-)$, $W^+ - W^-$, or W^+ itself. In each case, W has a specific distribution (depending on k_r) under the H_0 that the difference between the pairs follows a symmetric distribution around zero.

3.2.2 Dietterich 5 × 2 CV *t*-test

Dietterich (1998) proposed this test for determining whether there is a significant difference between the error rates of the two classifiers. Here, we show how to apply the test within the SRD framework for comparing the rankings of two models (π_1, π_2).

We start by taking random subsets $(A_1, \dots, A_k)$ and their complements $(A_1^c, \dots, A_k^c)$ from the n rows (objects) of sizes $\lfloor \frac{n}{2} \rfloor$ and $\lceil \frac{n}{2} \rceil$. In the original paper and in most applications, $k = 5$. Let $\underline{SRD}_{j,i}$ denote once again the $\underline{SRD}$ (from the reference) on fold A_i for the ranking π_j, and let $\underline{SRD}^c_{j,i}$ denote the same on the complement A_i^c.

First, look only at the fold (A_i, A_i^c). Calculate the differences between the normalized SRDs in both subsets,

$$\Delta_i := \underline{SRD}_{1,i} - \underline{SRD}_{2,i} \qquad \text{and} \qquad \Delta_{i^c} := \underline{SRD}^c_{1,i} - \underline{SRD}^c_{2,i},$$

and then calculate their average and sample variance:

$$\overline{\Delta_i} = \frac{1}{2}\left(\Delta_i + \Delta_{i^c}\right), \qquad s^2_{\Delta,i} = \left(\Delta_i - \overline{\Delta_i}\right)^2 + \left(\Delta_{i^c} - \overline{\Delta_i}\right)^2.$$

For large enough n, Δ_i and Δ_{i^c} are asymptotically normally distributed under the H_0 of uniformly selecting a ranking from S_n, thus we can get an approximately t-distributed test statistic:

$$\frac{\Delta_i}{\sqrt{\frac{1}{k}\sum_{i=1}^k s^2_{\Delta,i}}} \sim t_k,$$

and the same applies to the complementing fold. This way, we have $2k$ different test statistics, of which we can choose any for the evaluation of the hypothesis test.

3.2.3 Alpaydin 5×2 CV $\mathcal{F}$-test

The 5×2 CV $\mathcal{F}$-test of Alpaydin (1999) was proposed as an improvement on the CV t-test of Dietterich (1998). Here, we show how to apply the test within the SRD framework for comparing the rankings of two models (π_1, π_2). Up until the calculation of the Δ_i, Δ_{i^c} pairs and their sample variances for each fold, the setup is the same as in Sect. 3.2.2. However, the test statistic is approximately $\mathcal{F}$-distributed:

$$\frac{\frac{1}{2}\left(\Delta_i^2 + \Delta_{i^c}^2\right)}{s^2_{\Delta,i}} \sim \mathcal{F}_{2,1}.$$

Aggregating these for all folds results in:

$$\frac{\frac{1}{2k}\sum_{i=1}^k \left(\Delta_i^2 + \Delta_{i^c}^2\right)}{\frac{1}{k}\sum_{i=1}^k s^2_{\Delta,i}} \sim \mathcal{F}_{2k,k}.$$

Note that this assumes the independence of the calculated scores on the different folds, and again, the independence of the numerator and denominator too. Instead of aggregating the statistics of the folds, one can aggregate the dependent p-values coming from the folds separately. The correct aggregation of p-values would only assume the independence of the numerator and denominator.

4 Evaluation

How can the best-fitting option be selected in a complex decision situation where all the competing solutions seem fair in some way? One way is characterization—we break down performance into sub-cases. We come up with scenarios that the tests will likely face in practice, and then observe through a simulation, which test is more apt in which scenario. This type of analysis is especially suitable for rankings, as the differences between two rankings can be characterized fairly well.

We have evaluated three hybrid tests (Wilcoxon, Dietterich, and Alpaydin) under various parametrizations. The aim was to find the one that is the most efficacious in categorizing solutions. We have used the number of folds as parameters and varied it between 5 and 10. This interval is recommended by Hastie et al (2009, Chapter 7.10, p. 243) as a good compromise between the bias-variance trade-off.

4.1 Scenarios

We analyzed nine scenarios under three assumptions on the size of the rankings ($n = 7$, $n = 13$, and $n = 32$). The size options aim to represent the typical data sizes. Note that $n = 7$ is an odd number and almost too small for CV. However, since practitioners will not refrain from applying cross-validation for suboptimal data sizes, we felt the need to test this case as well. For $n = 7$ and $k = 7$, the Wilcoxon reduces to leave-one-out CV. If the number of folds, k, exceeds seven, we are forced to use bootstrapping; some rows are left out more than once.

All rankings that come from the same distribution are alike, rankings that come from different distributions are different in their own way.[1] Hence, we have looked at six scenarios for type II errors but only three for type I errors. These scenarios cover all typical situations, and other scenarios are unlikely to yield a new aspect as they would constitute a transition in between these.

In the following list, we have described the scenarios in detail. We started with a reference ranking, which is just the ordered list of numbers from 1 to n. Then, we compose two additional rankings, denoted by A and B, by making some transformations to the reference ranking. The null-hypothesis is that A and B come from the same distribution. Type I error occurs when a null hypothesis that is actually true is incorrectly rejected. Type II error occurs when a null hypothesis that is actually false is incorrectly accepted. For checking the type I error, we drew both rankings from the same distribution (Scenarios 1–3). Since the null-hypothesis is true we expect low rejection rates from the tests. In type II scenarios, the rankings were constructed in different ways (Scenarios 4–9). Since the null-hypothesis is false we expect high rejection rates from the tests. Table 1 displays the ranking transformations that were used to produce the rankings, and Fig. 1 shows their average distance from the reference.

[1] We are slightly paraphrasing Lev Tolstoy's Anna Karenina here.

Table 1 Ranking transformations used in the scenarios. We investigated the $n = 7,\ 13,\ 32$ cases

Identifier	Description
x	We apply x number of random inversions (switching of neighboring elements) on the reference ranking
xt	We apply x number of random inversions on the top $\lfloor n/2 \rfloor$ positions of the reference ranking
xb	We apply x number of random inversions on the bottom $\lceil n/2 \rceil$ positions of the reference ranking
1u	We select a random element (the underdog) from the bottom $\lceil n/2 \rceil$ positions of the reference ranking and switch it with the first element
4 m	We consecutively select four random positions between 1 and $n - \lceil n/4 \rceil$ from the reference ranking and switch the selected element at position s with the element at position $s + \lceil n/4 \rceil$

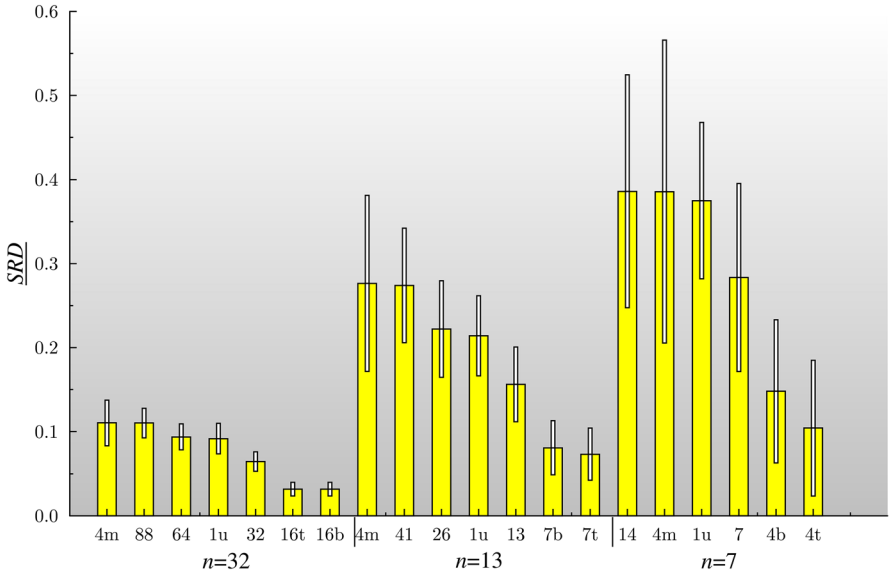

Fig. 1 Transformed rankings average distance from the reference ± standard deviation in normalized SRD

A scenario is defined as a pair $(a|b)$, where a and b refer to the transformations used to create rankings A and B respectively. For the definition of the transformations see Table 1.

1. $(2n|2n)$: This scenario investigates what happens if both rankings are drawn from the same distribution and their distance to the reference ranking is relatively large.
2. $(n|n)$: Similar to the previous one but with fewer inversions, so the rankings are closer to the reference.
3. RT I.: We picked a transformation uniformly and randomly from the set $2n$, n, $(n/2)t$, $(n/2)b$, $1u$, $4m$. Both rankings A and B were drawn from the

selected distribution. This scenario demonstrates what can we expect from a CV method regarding the type I error when we do not have prior information on the rankings' distribution.

4. $(2n|n/2)$: This scenario investigates what happens if the rankings are drawn from different distributions, that is, if their mean distance to the reference ranking is different, because the second ranking is a result of fewer inversions, so it is closer to the reference.
5. $(n|n/2)$: A noisier variant of the previous scenario. Here the average distance between rankings is smaller, hence, it is more difficult to distinguish between the rankings.
6. $((n/2)t|(n/2)b)$: This scenario shows what happens when the data is structured. For instance, for data collected from two periods, the solutions perform differently in the first and second periods. Note that the expected distance from the reference is approximately the same for both A and B (*cf.* Fig. 1).
7. $(2n|1u)$: This scenario tests the presence of outliers. Rankings A and B are of the same distance from the reference. However, the former is constructed by applying many small inversions, while in ranking B, we only swap one pair of elements. To illustrate this scenario, let us borrow an example from sports. Ranking A shows how the actual result of a sporting event, *e.g.* the soccer World Cup, differs from the preliminary ranking. Ranking B is the same as the preliminary ranking except that Burkina Faso wins, relegating Brazil to the second half of the points table. The preferability of ranking A or B depends on the application, but a CV method should be able to distinguish between the two rankings.
8. $(x|4m)$ Similar to the previous scenario, but having more, albeit less extreme, outliers. The number of inversions for ranking A, x, is chosen in a way that the expected distance from the reference for both rankings is approximately the same.
9. RT II.: We picked a transformation uniformly and randomly from the set $2n$, n, $(n/2)t$, $(n/2)b$, $1u$, $4m$. Ranking A was drawn from the selected distribution, and for B, we picked another transformation randomly. This scenario demonstrates what can we expect from a CV method regarding type II error when we do not have prior information on the rankings' distribution.

Figure 2 shows the relative frequency of distances between a perturbed ranking and the reference ranking in the $n = 32$ case. As the number of inversions grows the discrete distribution of the distance values becomes asymptotically normal (Diaconis and Graham 1977).

Manhattan distance satisfies the triangle inequality, hence if one ranking is close to the reference and the other is not, then they fall far from each other. There is significant overlap among the distributions generated by the distances of 16 and 32 random inversions. Thus we cannot expect any CV method to tell them apart 100 % of the time. Note that just because two rankings are of the same distance from the reference it does not necessarily mean that they are close to each other. Only that, in such cases, we cannot eliminate the possibility, so it is difficult to quantify the ideal rejection rate.

In the 64|16 scenario, however, the distributions overlap only at their tails. There is a very low probability that the rankings are of a similar distance from the

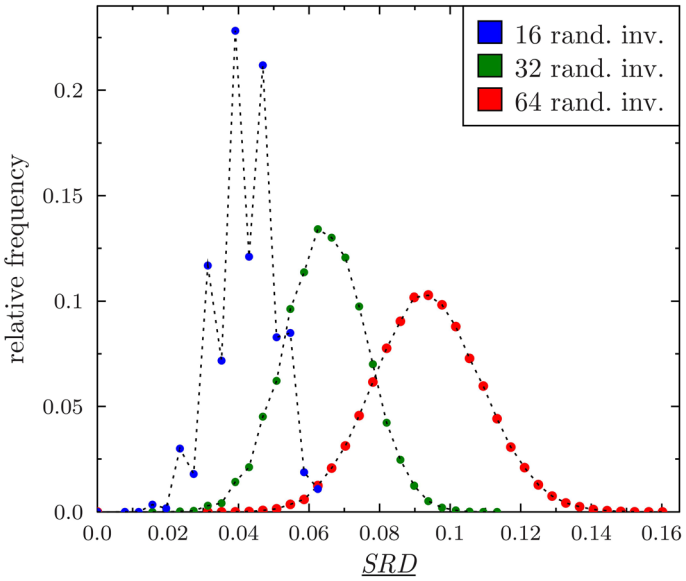

Fig. 2 The discrete distribution of the distances from reference under different transformations in the $n = 32$ case

reference, and even if they are, they may differ in completely disparate segments. CV methods should be able to distinguish between such rankings in the vast majority of the cases.

Let us note that the distribution depends on the choice of reference and the ranking generating scheme. In this paper, the reference is a fixed permutation and rankings are generated using inversions. There are alternative ways of generating random rankings, especially if the reference is a ranking that contains tied values.

4.2 Simulation results

For each scenario and data size, we performed 10 simulation rounds, each comprising 100,000 runs. We randomly generated instances of rankings A and B, applied the CV technique, and noted whether it rejected the null hypothesis at 5% level of significance (using the theoretical distribution corresponding to the respective statistical test). Tables 2, 3 and 4 show the average rate of rejections in the 10 rounds. In type I scenarios (1–3) the lower the rejection rate the better-the rankings come from the same distribution, so the CV method is expected to fail to reject the null hypothesis.[2] In type II scenarios (4–9), the opposite is true; the higher the rejection rate, the

[2] Note that, if all the rigorous assumptions of the applied statistical tests held firm, these numbers would be around 5%. Both Dietterich (1998) and Alpaydin (1999) knew well that their assumptions were not satisfied in general but dismissed concerns based on the empirical success of the tests.

Table 2 Null hypothesis testing for $n = 32$, rejection rate (%) in different scenarios

CV method / number of folds	type I. scenarios 64|64	32/32	RT I	type II. scenarios 64|16	32|16	16t|16b	64|1u	88|4m	RT II
Wilcoxon 5	35.5	30.3	37.5	95.1	61.1	16.8	51.5	53.0	74.1
Wilcoxon 6	35.7	30.3	37.8	95.9	62.6	15.6	51.8	53.0	75.2
Wilcoxon 7	43.7	37.5	42.9	97.3	69.1	21.5	57.0	62.8	81.2
Wilcoxon 8	47.2	40.8	45.9	97.9	72.2	24.0	60.4	65.8	83.1
Wilcoxon 9	55.9	50.3	53.2	98.9	79.7	32.3	66.5	73.0	86.3
Wilcoxon 10	59.4	53.5	54.9	99.1	81.2	36.5	67.8	76.4	87.3
Dietterich 5	3.0	2.7	2.6	22.5	6.8	0.6	2.0	3.0	11.6
Dietterich 6	3.2	3.0	2.8	24.3	7.5	0.7	1.5	3.2	12.3
Dietterich 7	3.3	3.0	3.0	25.2	7.7	0.7	1.2	3.3	12.6
Dietterich 8	3.5	3.2	3.2	26.2	8.2	0.8	1.0	3.5	13.0
Dietterich 9	3.6	3.3	3.4	26.7	8.3	0.8	0.8	3.7	13.2
Dietterich 10	3.8	3.4	3.5	27.2	8.5	0.9	0.7	3.8	13.5
Alpaydin 5	2.4	2.1	1.8	30.8	7.3	0.3	2.4	2.7	15.5
Alpaydin 6	2.6	2.1	1.9	36.8	8.5	0.3	2.1	2.8	18.1
Alpaydin 7	2.8	2.2	2.0	42.3	9.5	0.2	1.9	2.9	20.6
Alpaydin 8	3.0	2.4	2.2	47.0	10.7	0.2	1.8	3.1	22.9
Alpaydin 9	3.3	2.5	2.3	51.3	11.9	0.2	1.7	3.3	25.0
Alpaydin 10	3.5	2.7	2.5	54.9	13.0	0.2	1.7	3.5	26.8

Table 3 Null hypothesis testing for $n = 13$, rejection rate (%) in different scenarios

CV method / number of folds	type I. scenarios 26|26	13|13	RT I	type II. scenarios 26|7	13|7	7t|7b	26|1u	41|4m	RT II
Wilcoxon 5	30.4	25.3	33.4	64.6	33.7	13.1	39.7	45.0	59.9
Wilcoxon 6	26.9	21.8	30.3	61.7	30.1	10.1	36.0	41.3	57.0
Wilcoxon 7	37.7	31.5	39.6	72.5	40.8	16.9	45.3	55.7	68.0
Wilcoxon 8	36.3	30.0	38.1	71.4	39.4	15.7	43.9	54.4	67.2
Wilcoxon 9	46.4	40.5	47.8	79.8	50.3	23.9	52.2	64.2	73.7
Wilcoxon 10	48.5	41.7	48.0	80.6	51.2	26.0	52.8	66.4	74.5
Dietterich 5	3.7	3.5	3.1	9.9	4.6	0.7	2.3	4.4	7.7
Dietterich 6	4.1	3.9	3.6	10.8	5.1	0.8	2.1	4.8	8.3
Dietterich 7	4.5	4.2	4.0	11.4	5.5	0.9	2.1	5.1	8.7
Dietterich 8	4.7	4.5	4.4	11.9	5.8	1.0	2.1	5.4	9.1
Dietterich 9	5.0	4.7	4.6	12.3	6.2	1.2	2.0	5.7	9.4
Dietterich 10	5.2	4.9	4.8	12.7	6.3	1.3	2.1	5.8	9.7
Alpaydin 5	2.8	2.5	2.2	10.7	3.8	0.3	1.9	4.3	8.6
Alpaydin 6	3.2	2.7	2.3	12.7	4.3	0.2	1.8	4.9	10.0
Alpaydin 7	3.4	2.9	2.5	14.5	4.7	0.2	1.8	5.4	11.4
Alpaydin 8	3.8	3.2	2.8	16.5	5.3	0.1	1.8	5.9	12.7
Alpaydin 9	4.1	3.5	3.0	18.2	5.8	0.1	1.9	6.3	14.0
Alpaydin 10	4.5	3.8	3.2	19.9	6.3	0.1	1.9	6.8	15.3

Table 4 Null hypothesis testing for $n = 7$, rejection rate (%) in different scenarios

CV method / number of folds	type I. scenarios 14|14	7|7	RT I	type II. scenarios 14|4	7|4	4t|4b	14|1u	14|4m	RT II
Wilcoxon 5	19.5	15.4	18.1	31.2	16.4	9.6	25.4	24.4	35.8
Wilcoxon 6	16.3	12.6	15.1	27.6	13.5	6.8	21.8	21.2	32.1
Wilcoxon 7	29.4	24.4	24.2	44.8	26.2	13.0	35.9	34.5	49.7
Wilcoxon 8	36.4	31.5	33.2	51.2	31.9	22.1	41.7	41.0	55.0
Wilcoxon 9	44.9	40.2	43.6	59.3	39.6	32.4	48.7	48.2	61.6
Wilcoxon 10	44.2	39.3	42.2	58.5	38.4	30.8	47.8	47.6	60.5
Dietterich 5	5.2	5.4	7.9	6.1	2.5	3.3	5.4	6.8	4.1
Dietterich 6	6.0	6.4	9.0	7.1	3.2	3.7	6.1	7.9	5.0
Dietterich 7	6.4	6.9	9.5	7.6	3.6	4.1	6.7	8.7	5.6
Dietterich 8	6.9	7.3	10.0	8.1	4.1	4.5	7.1	9.3	6.2
Dietterich 9	7.2	7.7	10.4	8.4	4.3	4.7	7.4	9.8	6.5
Dietterich 10	7.5	7.9	10.7	8.7	4.7	5.0	7.7	10.2	7.0
Alpaydin 5	2.6	2.3	1.6	4.9	2.6	0.2	1.4	3.3	4.3
Alpaydin 6	2.9	2.4	1.8	5.6	2.9	0.2	1.5	3.9	5.3
Alpaydin 7	3.2	2.6	1.9	6.3	3.1	0.2	1.6	4.4	5.9
Alpaydin 8	3.6	2.9	2.2	7.2	3.6	0.2	1.7	5.0	6.9
Alpaydin 9	4.1	3.4	2.6	8.2	4.1	0.3	1.9	5.6	7.9
Alpaydin 10	4.5	3.6	2.7	8.9	4.4	0.3	2.1	6.1	8.8

better. We also observed the standard error on the average rejection rate of the 10 rounds, which falls under 0.002 in all scenarios.

There are several conclusions that we can infer from Tables 2, 3 and 4:

- No CV method excels in both type I and II situations.
- The Dietterich and Alpaydin tests rarely reject the null hypothesis. Thus, they excel in type I scenarios but perform poorly in type II scenarios.
- The opposite is true for Wilcoxon: it shines in type II situations but in type I cases the error rate is nowhere near the theoretical 5%.
- With some rare exceptions, increasing the number of folds raises the rate of rejection for all methods. Hence, there is a trade-off between the efficiency of methods in type I and type II scenarios.

As the tables show, different techniques excel in different scenarios. If practitioners have some preliminary knowledge about the distribution of the rankings that the investigated methods produce, they can choose the appropriate CV method. In many cases, however, these distributions are unknown. Therefore, to select the most suitable method for generic purposes, we need to dig deeper into the data.

4.3 Selection criteria

Here, we have listed some aspects according to which the data can be assessed. We have ranked the methods according to each aspect, and then aggregated the rankings using the Borda count.

Discriminative power (DISC): We have taken the absolute difference between the rejection rate in the RT II and RT I scenarios ($|\text{RT I} - \text{RT II}|$); the underlying idea is that the number of folds largely explains the rejection rate of the methods. The amount a method rejects RT II instances more often than RT I instances shows how well it can distinguish the two situations. We had no prior knowledge about the distribution of the rankings, hence we used the random scenarios. The larger the discriminative power the better.

Maximum distance from the best option (MAXDIFF): In each scenario, we identified the best option, that is, the CV method with the best rejection rate (lowest for type I/highest for type II scenarios). Then, for each method, we measured the difference between the method's performance in the scenario and the performance of the best option. Finally, a method was evaluated by the maximum of the differences across all scenarios. Small values indicate good performance.

Average distance from the best option (AVGDIFF): Same as the previous, with the exception that we calculated the average difference (instead of the maximum) from the best option. Again small values are preferred.

Balancedness (BLNC): We measured the absolute difference of type I and type II errors using the RT I and RT II columns (|RT I − (1 − RT II)|). The idea behind this is that we want to balance the errors—the smaller the difference, the better. Balancing the two types of error rates is a common method in biometrics, called crossover (or equal) error rate, see *e.g.* Ref. (Conrad et al. 2017, Chapter5). The smaller, the better.

Sum of Ranking Differences (SRD): We took the transpose of Tables 2, 3 and 4, hence the CV methods correspond to the columns (solutions), and the scenarios to the rows (objects). Reference is the row minimum in type I scenarios (the lowest rejection rate) and the row maximum in type II scenarios (the highest rejection rate). SRD calculates how far each CV method falls from the reference. Small SRD values indicate that the method is close to the desired reference.

Pair-wise correlation methods (CEPWAVG/WTPWAVG): Similarly to SRD, these methods work with the transposed data matrix (Héberger and Rajkó 2002). Two solutions (X1 and X2) were selected and checked to determine whether they were related to the reference (here the average), and whether both of their difference were positive (A); one of them was positive and the other was negative (B), or *vice versa* (C). The frequencies were counted for all possible pairs of scenarios. Then, two statistical tests—the conditional exact Fisher's test (CE), and Williams' t-test (Wt)—decide whether the frequencies of events B and C are significantly different or not; *i.e.,* one solution (say X1) is overriding X2, conversely, X1 loses against X2, or no decision can be made (tie). After that, the solutions were compared pairwise with the reference, considering all possible combinations (W05, W06,..., and A10). The solutions were further ranked according to the number of wins minus the number of losses, but the ranking was adjusted in the present case: probability-weighted ranking (pW) was used, *i.e.*, based on p(wins)-p(losses) scores. For both methods, larger values indicate good performance.

There are other possible aspects in terms of which the CV methods can be compared. However, since the analysis already contains seven different decision criteria, a new aspect has a small chance of turning over the aggregated ranking. For the aggregation method, we choose the Borda count. One advantage of this technique is its conceptual simplicity, but it fares well from a machine learning perspective as well (Burka et al. 2022). As Tables 5, 6 and 7 show, the results are fortunately rather straightforward.

4.4 Model validation

In this section, we show that the proposed scenarios are meaningful and grasp the behavior of real data. For demonstration, we examined two unrelated datasets: laboratory performances in a quality control program to test type I situations and Elo-scores from a chess championship to test type II situations.

In addition, we conducted tests against rankings generated from the Mallows model using various parameters. While the data is synthetically generated through

Table 5 Ranking of CV-methods in the $n = 32$ case

	DISCPOW	MAXDIFF	AVGDIFF	BLNC	SRD	CEPWAVG	WTPWAVG	Borda
Wilcoxon 5	4	1	6	1	3.5	5.5	2	103
Wilcoxon 6	2	2	5	2	3.5	5.5	3	103
Wilcoxon 7	1	3	4	3	3.5	2.5	6	103
Wilcoxon 8	3	4	3	4	3.5	2.5	11	95
Wilcoxon 9	5	5	2	5	3.5	2.5	16	87
Wilcoxon 10	6	6	1	6	3.5	2.5	17	84
Dietterich 5	18	18	18	18	15.5	16	1	21.5
Dietterich 6	17	17	17	17	15.5	16	4	22.5
Dietterich 7	16	16	16	16	15.5	10.5	7	29.0
Dietterich 8	15	15	15	15	15.5	16	15	19.5
Dietterich 9	14	14	14	14	10.5	9	13	37.5
Dietterich 10	13	12	13	13	10.5	10.5	18	36.0
Alpaydin 5	12	13	12	12	7.5	7.5	5	57
Alpaydin 6	11	11	11	11	7.5	7.5	8	59
Alpaydin 7	10	10	10	10	10.5	12.5	14	49
Alpaydin 8	9	9	9	9	10.5	12.5	12	55
Alpaydin 9	8	8	8	8	15.5	16	9	53.5
Alpaydin 10	7	7	7	7	15.5	16	10	56.5

Table 6 Ranking of CV-methods in the $n = 13$ case

	DISCPOW	MAXDIFF	AVGDIFF	BLNC	SRD	CEPWAVG	WTPWAVG	Borda
Wilcoxon 5	4	2	5	2	3	2.5	4	103.5
Wilcoxon 6	3	1	6	4	6	6	5	95.0
Wilcoxon 7	2	4	3	3	3	2.5	1	107.5
Wilcoxon 8	1	3	4	1	3	2.5	2	109.5
Wilcoxon 9	6	5	2	5	3	2.5	3	99.5
Wilcoxon 10	5	6	1	6	3	5	6	94.0
Dietterich 5	18	18	18	17	15.5	15.5	9	15
Dietterich 6	17	16	17	16	15.5	15.5	10	19
Dietterich 7	16	15	16	14	15.5	15.5	12	22
Dietterich 8	15	14	15	13	15.5	15.5	14	24
Dietterich 9	14	13	13	11	15.5	15.5	16	28
Dietterich 10	13	12	12	10	15.5	15.5	18	30
Alpaydin 5	12	17	14	18	9.5	9.5	7	39
Alpaydin 6	11	11	11	15	9.5	9.5	8	51
Alpaydin 7	10	10	10	12	9.5	9.5	11	54
Alpaydin 8	9	9	9	9	9.5	9.5	13	58
Alpaydin 9	8	8	8	8	9.5	9.5	15	60
Alpaydin 10	7	7	7	7	9.5	9.5	17	62

Table 7 Ranking of CV-methods in the $n = 7$ case

	DISCPOW	MAXDIFF	AVGDIFF	BLNC	SRD	CEPWAVG	WTPWAVG	Borda
Wilcoxon 5	5	2	5	5	3	4	5	97.0
Wilcoxon 6	6	3	9	6	3	4	4	91.0
Wilcoxon 7	1	1	4	4	1	6	6	103.0
Wilcoxon 8	2	4	3	3	3	4	2	105.0
Wilcoxon 9	4	11	1	2	5.5	1.5	3	98.0
Wilcoxon 10	3	10	2	1	5.5	1.5	1	102.0
Dietterich 5	15	12	12	12	15.5	15.5	18	26.0
Dietterich 6	11	9	11	11	15.5	15.5	16	37.0
Dietterich 7	12	8	10	10	15.5	15.5	17	38.0
Dietterich 8	13	7	8	9	15.5	15.5	13	45.0
Dietterich 9	14	6	7	8	15.5	15.5	15	45.0
Dietterich 10	16	5	6	7	15.5	15.5	14	47.0
Alpaydin 5	18	18	18	18	9.5	8.5	11	25.0
Alpaydin 6	17	17	17	17	9.5	11.5	10	27.0
Alpaydin 7	10	16	16	16	9.5	8.5	8	42.0
Alpaydin 8	9	15	15	15	9.5	8.5	9	45.0
Alpaydin 9	8	14	14	14	9.5	11.5	12	43.0
Alpaydin 10	7	13	13	13	9.5	8.5	7	55.0

Table 8 Cross-validation results for the OIL dataset. The measurements of four laboratories (L1, L4, L10, L11) were compared. Each cell shows the average rejection rate (%) in a sample of 100,000 runs. The last column shows the average of the pairwise comparisons

	L1-L4	L1-L10	L1-L11	L4-L10	L4-L11	L10-L11	Avg
Wilcoxon 5	3.6	2.6	3.6	8.0	0.5	8.8	4.5
Wilcoxon 6	2.7	1.6	2.7	9.5	0.1	9.8	4.4
Wilcoxon 7	1.6	2.0	1.6	7.0	0.0	9.3	3.6
Wilcoxon 8	2.4	4.1	2.4	12.5	0.0	18.4	6.6
Wilcoxon 9	8.5	5.0	8.5	24.3	0.0	25.8	12.0
Wilcoxon 10	6.1	7.2	6.2	26.4	0.0	31.5	12.9
Dietterich 5	1.1	1.1	0.2	0.2	1.1	0.3	0.7
Dietterich 6	1.7	1.7	0.3	0.3	1.8	0.3	1.0
Dietterich 7	1.1	1.1	0.2	0.2	1.2	0.4	0.7
Dietterich 8	1.5	1.6	0.3	0.3	1.7	0.4	0.9
Dietterich 9	1.3	1.3	0.2	0.2	1.4	0.4	0.8
Dietterich 10	1.5	1.5	0.2	0.2	1.7	0.4	0.9
Alpaydin 5	0.8	0.1	0.8	0.0	0.7	0.0	0.4
Alpaydin 6	0.5	0.0	0.5	0.0	0.5	0.0	0.3
Alpaydin 7	0.3	0.0	0.3	0.0	0.3	0.0	0.2
Alpaydin 8	0.2	0.0	0.2	0.0	0.2	0.0	0.1
Alpaydin 9	0.3	0.0	0.3	0.0	0.3	0.0	0.1
Alpaydin 10	0.2	0.0	0.2	0.0	0.1	0.0	0.1

simulation, this approach serves as a suitable validation method, given that the generating process differs from the one employed in our scenarios.

4.4.1 OIL dataset

To verify the confidence of their analytical methods, laboratories participate in a comparison program, where they have to determine some characteristics of a homogeneous sample under documented conditions. In our example, Polycyclic Aromatic Hydrocarbon contents in 16 edible oil samples (OIL) were reported by each participating laboratory (Škrbić et al. 2013). Since the laboratories work with the same substances, the expectation is that they report the same values within a small statistical error. Reference values were provided by the European Union Reference Laboratory for PAHs in food (EU-RL-PAH).

From the 15 laboratories, we selected four that have shown very little discrepancy from the reference. Indeed, the various statistical tests applied in Ref. Škrbić et al. (2013) failed to show any significant difference in the measured values of these four laboratories. Thus, both expectation and empirical evidence point toward that the reported measures come from the same distribution. Clearly, this is a type I situation, where CV methods should fail to reject the null hypothesis.

CV has a stochastic element, in each run different rows are selected. Thus, rerunning a CV method on the same data may yield different results. To uncover the characteristic rejection rate on the OIL dataset, we ran each CV method 100 000 times for each pair of laboratories. Table 8 compiles the resulting rejection rates.

Sixteen substances are compared in the OIL dataset, with the previous notation, $n = 16$. Given the small differences between the measured values, it is as if

Table 9 Cross-validation results for the CHESS dataset. Preliminary Elo ratings are cross-tested with tournament performances. Each cell shows the average rejection rate (%) in a sample of 100 000 runs

	prelim. *vs.* tour. perf.		prelim. *vs.* tour. perf.
Wilcoxon 5	99.99	Dietterich 8	18.31
Wilcoxon 6	100.00	Dietterich 9	18.79
Wilcoxon 7	100.00	Dietterich 10	19.01
Wilcoxon 8	100.00	Alpaydin 5	18.54
Wilcoxon 9	100.00	Alpaydin 6	22.64
Wilcoxon 10	100.00	Alpaydin 7	27.24
Dietterich 5	15.74	Alpaydin 8	31.99
Dietterich 6	16.91	Alpaydin 9	37.15
Dietterich 7	17.68	Alpaydin 10	42.33

we applied 8 or 10 random inversions on the reference. The closest scenario in our analysis is 13|13 under the $n = 13$ case. The result (Avg column in Table 8) indeed resembles what we see in Table 3. All methods managed to improve their efficiency, most notably Wilcoxon 5 to 8 is very close to the desired 5% threshold. In comparison, the rejection rate of the Wilcoxon variants ranged between 25% and 40% during the simulations under the artificial scenario.

4.4.2 Chess dataset

There is a long tradition in chess to measure the performance of the players. The Elo score system quantifies the playing strengths. The difference between Elo ratings can be translated into winning probabilities: how likely is it that the stronger player (with a higher rating) beats the weaker one (with a lower rating)?

We chose the Grand Swiss tournament of 2019, and looked at the preliminary Elo ratings, tournament performance (also measured in Elo points), and post-tournament Elo ratings of the top 32 players.[3]

Tournament performance is calculated based on the opponents' Elo rating and match scores and it may depend on various factors, including quality of preparation and match pairings, but also on unexpected events like the mood of the players or a sudden illness. Based on the wins and losses, the post-tournament Elo rating differs from the preliminary rating by a couple of Elo points. Tournament performance, on the other hand, fluctuates wildly.

Sziklai et al. (2022) compares the performances of 900 players in six different tournaments and derives an empirical distribution of the Elo rating difference. The distribution resembles a Gaussian curve, but with a fat tail, meaning that extremely good or bad performances are not that uncommon.

The same is true for our data. The ranking of the 32 players based on their performances visibly differs from either the preliminary or the post-tournament rankings. Preliminary rankings are based on the match history of the players. The preliminary

[3] Data was gathered from ChessResults.com: http://chess-results.com/tnr478041.aspx?lan=1 &art=1 & flag=30.

Elo scores might not be up-to-date since it could have been months since the last rated game was played by the contestants. Although as we mentioned, there are a plethora of reasons why a chess match is decided one way or the other, most wins are based on merit. Arguably, the post-tournament ratings are the closest to the true power ranking of the players, hence we choose it as the reference. From this point of view, preliminary ratings and tournament performances are two perturbations of different amplitude. This is clearly a type II situation, CV methods should be able to distinguish between the two data columns.

As in the previous case, we ran 100,000 simulations to uncover the characteristic rejection rates of the methods. Table 9 compiles the result. Since the number of players is the same as the number of rows in our first simulation setup, the results are directly comparable to the numbers in Table 2. However, the difference between the rankings is much larger. The most powerful transformation we applied in the $n = 32$ case is 88 inversions which created on average a 0.11 normalized SRD score (*cf.* Fig. 1). Here, the normalized SRD distance between the preliminary ranking and the reference is 0.15, while the tournament performance differs with a hefty 0.29 SRD score. This means that the difference between the rankings is more pronounced. Still, we would expect something similar to the 64|16 scenario and indeed that is what we see in Table 9. Again, the only difference is that the Wilcoxon variants perform better (with 100 % efficacy!) than in the simulation.

To conclude, the test runs on the real data show very similar results to the simulations of the appropriate scenarios. The slight differences stem from the variation in the number of rows or the magnitude of the perturbations. By and large, the scenarios paint an accurate picture of what is expected under various circumstances.

4.4.3 Mallows model

The Mallows model is used in the fields of marketing and political science for preference learning and choice modeling (Vitelli et al. 2017). To generate rankings from the Mallows model, we used the BayesMallows package (Sørensen et al. 2020). In the Mallows model, the probability density for ranking data depends on a consensus ranking (the reference) and a scale parameter,[4] denoted by α_0. The higher the parameter, the closer the generated ranking to the reference.

We selected α_0 values that were comparable in terms of distance to those of the previously tested scenarios. As Fig. 3 shows, the standard deviations are somewhat larger. Thus, there is a greater variability in the generated rankings, which the methods do not handle equally well.

For type I. testing both rankings were generated using the same parameter, for type II. cases different parameters were chosen. The rejection rates are compiled in Table 10. The results are generally consistent with our previous findings, though there are two peculiarities.

[4] Different authors refer to the scale parameter with different Greek letters. Here we use the notation of the BayesMallows package.

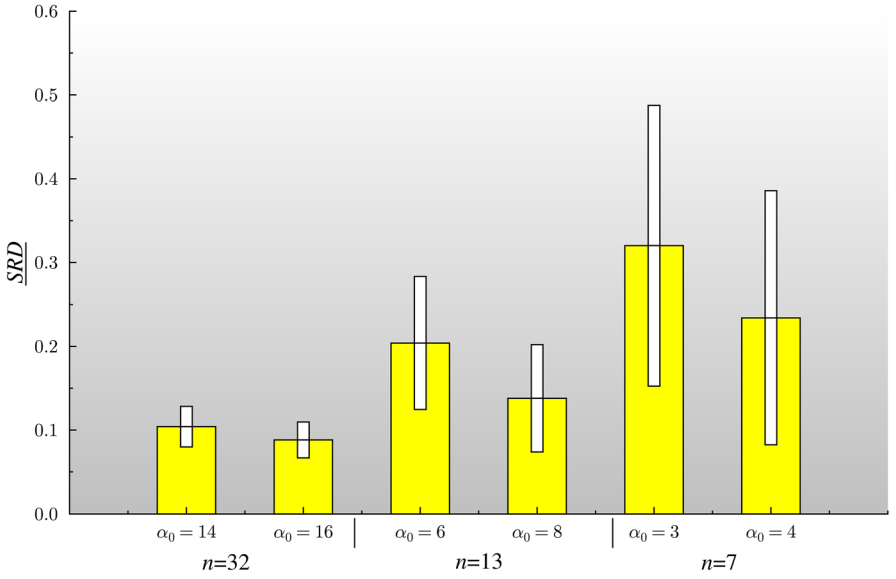

Fig. 3 Mallows rankings average distance from the reference ± standard deviation in normalized SRD

Table 10 Null hypothesis testing for the Mallows model, rejection rate (%) in different scenarios

CV method / number of folds	type I. scenarios			type II. scenarios		
	n=32 $\alpha_0 = 16\|16$	n=13 $\alpha_0 = 8\|8$	n=7 $\alpha_0 = 4\|4$	n=32 $\alpha_0 = 16\|14$	n=13 $\alpha_0 = 8\|6$	n=7 $\alpha_0 = 4\|3$
Wilcoxon 5	47.1	39.8	26.3	52.0	49.0	30.1
Wilcoxon 6	47.3	36.3	22.7	52.4	45.8	26.6
Wilcoxon 7	55.4	47.0	38.3	60.3	56.5	42.3
Wilcoxon 8	58.4	45.5	37.9	63.3	54.9	42.2
Wilcoxon 9	65.9	55.9	38.5	70.3	64.0	42.0
Wilcoxon 10	69.1	57.1	37.8	73.0	65.9	42.1
Dietterich 5	3.5	4.3	5.8	4.5	5.8	6.4
Dietterich 6	3.9	4.7	6.9	4.8	6.3	7.3
Dietterich 7	4.1	5.2	7.6	5.2	6.8	7.9
Dietterich 8	4.3	5.3	8.0	5.3	7.2	8.5
Dietterich 9	4.4	5.5	8.4	5.5	7.4	9.0
Dietterich 10	4.5	5.9	8.9	5.7	7.7	9.3
Alpaydin 5	3.6	3.6	2.7	4.9	5.8	3.7
Alpaydin 6	4.1	4.3	3.1	5.8	6.9	4.3
Alpaydin 7	4.5	4.8	3.8	6.5	7.9	4.9
Alpaydin 8	5.1	5.5	4.2	7.2	8.9	5.7
Alpaydin 9	5.6	6.1	4.7	7.9	9.6	6.2
Alpaydin 10	6.0	6.6	5.2	8.8	10.8	7.0

The Wilcoxon method performs slightly worse, while the Dietterich and Alpaydin are slightly better than in the original simulation. Another interesting phenomenon is the monotonicity violations. Wilcoxon with 6 folds shows lower rejection rates in some scenarios than Wilcoxon with 5 folds and the same is true for Wilcoxon with 8 and 7 folds. These help in type I. scenarios, but hinder the method in type II. cases. Given that Wilcoxon encounters greater challenges with the former, this further

supports our conclusion that Wilcoxon with 8 folds is the most viable (or rather, the least problematic) choice overall.

5 Discussion

With some rare exceptions, all selection criteria rank the six Wilcoxon variants in the first six places. The downside of Wilcoxon is that it is too sensitive: it picks up on subtle differences in the data and rejects the null hypothesis even if it ought to not. The case studies show that the simulation scenarios adequately model what happens on real data. The only difference is that in both case studies Wilcoxon's performance exceeded expectations.

Dietterich and Alpaydin are near the desired 5% threshold in type I situations. In type II situations Alpaydin outperforms Dietterich, although it is still very bad compared to Wilcoxon. Hence, if the practitioner is not afraid of categorizing different things as the same they may select Alpaydin CV. Among the Alpaydin variants, the 10-fold version performs the best.[5] Although it is uncommon to apply a fold greater than 10, it may improve the test's performance in type II situations. The simulation results also suggest that Alpaydin's power becomes better as we increase the data size, that is, the number of rows in the analysis.

That being said, it is rarely the case that we only care about one type of error. Both Alpaydin and Dietterich perform poorly in type II situations. How poorly?

Figure 2 sheds light on the expected rejection rates in the 64|16 and 32|16 scenarios. Although it is difficult to quantify the ideal rejection rates, in the 64|16 scenario a rejection rate of 90+% is definitely preferable to a rate of 20–50%. Even in the 32|16 scenario, a rate of 60–80% is closer to the truth than a rate of 5–15%.

Another problem with Dietterich and Alpaydin is that they are incapable of recognizing 2n|1u or x|4 m type of scenarios, although such structures in the data can be easily detected by the naked eye. Interestingly, Wilcoxon performs convincingly in these scenarios. One possible reason behind the difference in the performances can be the way these CV methods leave out rows. Dietterich and Alpaydin leave out half the rows in each fold, while Wilcoxon only excludes 10–20%.

It is somewhat understandable that all CV methods struggle with the top *vs.* bottom $((n/2)t|(n/2)b)$ type of scenarios. Row selection is randomized for all three CV methods, thus they cannot distinguish between the first and second part of the data. Real datasets, however, are often composed of blocks so this scenario is very much relevant. The only remedy is if we encompass such lessons in the CV process and let the row selection follow some pattern instead of randomness. Note that Wilcoxon outperformed the other two methods in this aspect as well.

Let us summarize our findings.

[5] For large n, Borda seems to favor Alpaydin 5–6 over 10 (see Table 5). However, this happens because SRD and the pairwise correlation methods do not consider absolute differences of the underlying data. Alpaydin 5 and 6 are indeed slightly better in 64|1u and 88|4 m situations, but Alpaydin 10 is much better in 64|16 cases.

- Based on the aggregated indicators, Wilcoxon's test combined with eightfold cross-validation (in short Wilcoxon 8) performed the best, although Wilcoxon 7 was also very good. While the simulations show that Wilcoxon's rejection rate in type I situations is subpar, on real datasets its performance is satisfactory.
- The Borda scores indicate that Alpaydin is to be preferred over Dietterich under all data sizes.
- Alpaydin prevails in type I situations, but only performs somewhat satisfactorily in type II scenarios if both the number of folds and the size of the data (n) are large.
- None of the methods were particularly good. Therefore, future research is needed to discover a more efficacious combination of a statistical test and a suitable CV method.

6 Summary and conclusions

We tested how successful the combinations of statistical tests and cross-validation methods are in distinguishing rankings that come from various distributions. Despite the widespread usage of rankings, to our knowledge, this is the first paper that has tackled this problem. Our method of analysis is innovative, in the sense that we devised simulation scenarios to uncover the strengths and weaknesses of certain methods.

We investigated three tests, Wilcoxon's, Dietterich's, and Alpaydin's coupled with cross-validation with folds ranging from 5 to 10, making 18 variants altogether (Tables 2, 3, 4). The simulation data was ranked according to various decision criteria, and the rankings were then aggregated to choose a winner. Wilcoxon test with eight folds proved to be the best. While Wilcoxon is a bit too sensitive in type I situations, it is the only method that performs well in type II scenarios and is especially good at picking up structures in the data. Despite Wilcoxon's type I rejection rates in the simulations being unsatisfactory, its performance on real data proved to be solid. The Dietterich and Alpaydin tests fared poorly in type II situations, although Alpaydin was somewhat better than Dietterich. Alpaydin should only be chosen if the practitioner does not want to err in type I situations. In such cases, Alpaydin 10 is recommended. Tests with real data affirmed our findings.

Another interesting observation is that Alpaydin 10 dominates the fivefold Alpaydin test in almost every aspect (Tables 5, 6, 7), even though in practice, the latter is used almost always. It would be interesting to see how Alpaydin behaves with an even higher fold number.

We compared cross-validation methods in the framework of Sum of Ranking Differences. That means, that one of the rankings was fixed as a reference and the differences between the rankings were measured by Manhattan distance or L_1-norm. There are other meaningful metrics, for instance, the number of inversions (Kendall tau) or the Spearman distance (squares of differences). An interesting future research direction is to test whether the choice of distance metric has any effect on the results—although we do not expect big surprises in this aspect.

Cross-validation is one way to assign uncertainties to rankings—for simple comparison, there are a variety of distance metrics available. Often, CV is applied when distance metrics show little or no difference. This cannot be readily resolved by applying a different measure since distance metrics usually correlate. Although we tested the overall performance of Wilcoxon, Alpaydin, and Dietterich's tests combined with CV, it would be interesting to see their efficiency when the rankings are generated in a way that their distance to the reference is fixed.

In the simulations, rankings were generated using random inversions starting from a fixed permutation. Later, during validation, the calculations were repeated with rankings drawn from the Mallows model. The results were largely consistent. However, on real data, the cross-validation methods demonstrated slightly improved performance. It would be intriguing to explore how the results vary across a wider range of ranking-generating schemes, especially in cases where the starting ranking is not a permutation but a ranking that contains tied values.

Finally, there is a lot of room for improvement regarding the rejection rates. Even the best cross-validation method was not particularly good—there were scenarios where it performed poorly. Future research is needed to devise a more efficacious CV method to distinguish different rankings. Beside the choice of statistical test, sampling might also play a role. When generating folds some of the removed rows might overlap. Instead of random sampling, a structured selection might enhance the efficiency of the tests.

Acknowledgements Balázs R. Sziklai is the grantee of the János Bolyai Research Scholarship of the Hungarian Academy of Sciences and is also supported by the ÚNKP-23-5 New National Excellence Program of the Ministry for Culture and Innovation from the source of the National Research, Development and Innovation Fund. This research was supported by the National Research, Development and Innovation Office of Hungary OTKA grants K 138945 (Sziklai) and K 134260 (Héberger).

Funding Open access funding provided by Corvinus University of Budapest.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Abonyi J, Ipkovich A, Dörgő G et al (2023) Matrix factorization-based multi-objective ranaking-what makes a good university? Plos One 18(4):1–30. https://doi.org/10.1371/journal.pone.0284078

Alaiz-Rodríguez R, Parnell AC (2020) An information theoretic approach to quantify the stability of feature selection and ranking algorithms. Knowl Based Syst 195:105745. https://doi.org/10.1016/j.knosys.2020.105745

Alpaydin E (1999) Combined 5×2 cv F Test for comparing supervised classification learning algorithms. Neural Comput 11:1885–1892. https://doi.org/10.1162/089976699300016007

Barlow GW, Ballin PJ (1976) Predicting and assessing dominance from size and coloration in the polychromatic midas cichlid. Anim Behav 24(4):793–813. https://doi.org/10.1016/S0003-3472(76)80010-3

Bartholdi J, Tovey CA, Trick MA (1989) Voting schemes for which it can be difficult to tell who won the election. Soc Choice Welf 6(2):157–165

Brandenburg FJ, Gleißner A, Hofmeier A (2013) Comparing and aggregating partial orders with Kendall tau distances. Discret Math Algorithms Appl 05(02):1360003. https://doi.org/10.1142/S1793830913600033

Burka D, Puppe C, Szepesváry L et al (2022) Voting: a machine learning approach. Eur J Oper Res 299(3):1003–1017. https://doi.org/10.1016/j.ejor.2021.10.005

Conrad E, Misenar S, Feldman J (2017) Chapter 5—Domain 5: Identity and access management (controlling access and managing identity). In: Conrad E, Misenar S, Feldman J (eds) Eleventh Hour CISSP® (Third Edition). Syngress, pp 117–134, https://doi.org/10.1016/B978-0-12-811248-9.00005-X

Crispino M, Mollica C, Astuti V et al (2023) Efficient and accurate inference for mixtures of mallows models with spearman distance. Stat Comput 33(5):98. https://doi.org/10.1007/s11222-023-10266-8

Diaconis P, Graham RL (1977) Spearman's footrule as a measure of disarray. J R Stat Soc Ser B (Methodological) 39(2):262–268. https://doi.org/10.1111/j.2517-6161.1977.tb01624.x

Dietterich TG (1998) Approximate statistical tests for comparing supervised classification learning algorithms. Neural Comput 10(7):1895–1923. https://doi.org/10.1162/089976698300017197

Falivene O, Cabrera L, Tolosana-Delgado R et al (2010) Interpolation algorithm ranking using cross-validation and the role of smoothing effect. A coal zone example. Comput Geosci 36(4):512–519. https://doi.org/10.1016/j.cageo.2009.09.015

Farshadfar E, Amiri R (2016) In vitro application of integrated selection index for screening drought tolerant genotypes in common wheat. Acta Agric Slovenica 107(2):335. https://doi.org/10.14720/aas.2016.107.2.07

Gere A, Rácz A, Bajusz D et al (2021) Multicriteria decision making for evergreen problems in food science by sum of ranking differences. Food Chem 344:128617. https://doi.org/10.1016/j.foodchem.2020.128617

Gere A, Szakál D, Héberger K (2022) Multiobject optimization of national football league drafts: comparison of teams and experts. Appl Sci. https://doi.org/10.3390/app12136303

Gyarmati L, Orbán-Mihálykó E, Mihálykó C et al (2023) Aggregated rankings of top leagues' football teams: application and comparison of different ranking methods. Appl Sci. https://doi.org/10.3390/app13074556

Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction, 2nd edn. Springer series in statistics. Springer, New York

Héberger K (2010) Sum of Ranking Differences compares methods or models fairly. TrAC Trends Anal Chem 29(1):101–109. https://doi.org/10.1016/j.trac.2009.09.009

Héberger K, Kollár-Hunek K (2011) Sum of Ranking Differences for method discrimination and its validation: comparison of ranks with random numbers. J Chemom 25(4):151–158. https://doi.org/10.1002/cem.1320

Héberger K, Rajkó R (2002) Generalization of pair correlation method (PCM) for non-parametric variable selection. J Chemom 16(8–10):436–443. https://doi.org/10.1002/cem.748

Hild M, Spohn W (2008) The measurement of ranks and the laws of iterated contraction. Artif Intell 172(10):1195–1218. https://doi.org/10.1016/j.artint.2008.03.002

Irurozki E, Calvo B, Lozano JA (2016) PerMallows: an R package for mallows and generalized mallows models. J Stat Softw 71(12):1–30

Jiang J, Ma Q, Jiang X et al (2021) Ranking list preservation for feature matching. Pattern Recognit 111:107665. https://doi.org/10.1016/j.patcog.2020.107665

Kollár-Hunek K, Héberger K (2013) Method and model comparison by sum of ranking differences in cases of repeated observations (ties). Chemom Intell Lab Syst 127:139–146. https://doi.org/10.1016/j.chemolab.2013.06.007

Kumar R, Vassilvitskii S (2010) Generalized distances between rankings. In: Proceedings of the 19th international conference on World Wide Web. Association for computing machinery, New York, USA, pp 571-580, https://doi.org/10.1145/1772690.1772749

Lee PH, Yu PLH (2010) Distance-based tree models for ranking data. Comput Stat Data Anal 54(6):1672–1682. https://doi.org/10.1016/j.csda.2010.01.027

Lee PH, Yu PLH (2012) Mixtures of weighted distance-based models for ranking data with applications in political studies. Comput Stat Data Anal 56(8):2486–2500. https://doi.org/10.1016/j.csda.2012.02.002

Lin S (2010) Rank aggregation methods. WIREs Comput Stat 2(5):555–570. https://doi.org/10.1002/wics.111

Lockwood J, Louis TA, McCaffrey DF (2002) Uncertainty in rank estimation: implications for value-added modeling accountability systems. J Educ Behav Stat 27(3):255–270. https://doi.org/10.3102/10769986027003255

Mallows CL (1957) Non-Null Ranking Models I. Biometrika 44(1/2):114–130

Mollica C, Tardella L (2014) Epitope profiling via mixture modeling of ranked data. Stat Med 33(21):3738–3758. https://doi.org/10.1002/sim.6224

Mollica C, Tardella L (2017) Bayesian Plackett-Luce mixture models for partially ranked data. Psychometrika 82(2):442–458. https://doi.org/10.1007/s11336-016-9530-0

Moorthy NHN, Kumar S, Poongavanam V (2017) Classification of carcinogenic and mutagenic properties using machine learning method. Comput Toxicol 3:33–43. https://doi.org/10.1016/j.comtox.2017.07.002

Negahban S, Oh S, Thekumparampil KK et al (2018) Learning from comparisons and choices. J Mach Learn Res 19(40):1–95

Orbán-Mihálykó É, Mihálykó C, Gyarmati L (2023) Evaluating the capacity of paired comparison methods to aggregate rankings of separate groups. Central Eur J Oper Res. https://doi.org/10.1007/s10100-023-00839-3

Palmer D, Höck B, Kimberley M et al (2009) Comparison of spatial prediction techniques for developing Pinus Radiata productivity surfaces across New Zealand. For Ecol Manag 258(9):2046–2055. https://doi.org/10.1016/j.foreco.2009.07.057

Plackett RL (1975) The analysis of permutations. J R Stat Soc Ser C (Appl Stat) 24(2):193–202

Qian Z, Yu PLH (2019) Weighted distance-based models for ranking data using the R package rankdist. J Stat Softw 90(5):1–31. https://doi.org/10.18637/jss.v090.i05

Rosander AC (1936) The standard error of a mean rank order. J Educ Psychol 27(3):193–196. https://doi.org/10.1037/h0057950

Škrbić B, Héberger K, Đurišić Mladenović N (2013) Comparison of multianalyte proficiency test results by Sum of Ranking Differences, principal component analysis, and hierarchical cluster analysis. Anal Bioanal Chem 405:8363–8375. https://doi.org/10.1007/s00216-013-7206-5

Sørensen Ø, Crispino M, Liu Q et al (2020) Bayesmallows: an R package for the Bayesian mallows model. R J 12(1):324–342. https://doi.org/10.32614/RJ-2020-026

Staudacher J, Sziklai BR, Olsson L et al (2023) rSRD: Sum of Ranking Differences statistical test. https://doi.org/10.32614/CRAN.package.rSRD, https://CRAN.R-project.org/package=rSRD, R package version 0.1.7

Švendová V, Schimek MG (2017) A novel method for estimating the common signals for consensus across multiple ranked lists. Comput Stat Data Anal 115:122–135. https://doi.org/10.1016/j.csda.2017.05.010

Sziklai BR, Héberger K (2020) Apportionment and districting by Sum of Ranking Differences. Plos One 15(3):e0229209. https://doi.org/10.1371/journal.pone.0229209

Sziklai BR, Biró P, Csató L (2022) The efficacy of tournament designs. Comput Oper Res. https://doi.org/10.1016/j.cor.2022.105821

Tavanaei A, Gottumukkalay R, Maida AS et al (2018) Unsupervised learning to rank aggregation using parameterized function optimization. In: 2018 International joint conference on neural networks (IJCNN). IEEE, Rio de Janeiro, pp 1–8, https://doi.org/10.1109/IJCNN.2018.8489160

Tehrani AF, Cheng Weiwei, Hullermeier E (2012) Preference learning using the Choquet integral: the case of multipartite ranking. IEEE Trans Fuzzy Syst 20(6):1102–1113. https://doi.org/10.1109/TFUZZ.2012.2196050

Triantafilis J, Odeh I, McBratney A (2001) Five geostatistical models to predict soil salinity from electromagnetic induction data across irrigated cotton. Soil Sci Soc Am J 65(3):869–878. https://doi.org/10.2136/sssaj2001.653869x

Vajna B, Farkas A, Pataki H et al (2012) Testing the performance of pure spectrum resolution from Raman hyperspectral images of differently manufactured pharmaceutical tablets. Anal Chim Acta 712:45–55. https://doi.org/10.1016/j.aca.2011.10.065

Vitelli V, Sørensen Ø, Crispino M et al (2017) Probabilistic preference learning with the Mallows rank model. J Mach Learn Res 18(1):5796–5844

Volkovs MN, Zemel RS (2014) New learning methods for supervised and unsupervised preference aggregation. J Mach Learn Res 15(1):1135–1176

West C (2018) Statistics for analysts who hate statistics, Part VII: sum of ranking differences (SRD)s. LCGC N Am 36:2–6

Wilcoxon F (1945) Individual comparisons by ranking methods. Biomet Bull 1(6):80. https://doi.org/10.2307/3001968

Xu H, Alvo M, Yu PLH (2018) Angle-based models for ranking data. Comput Stat Data Anal 121:113–136. https://doi.org/10.1016/j.csda.2017.12.004

Yu PLH, Gu J, Xu H (2019) Analysis of ranking data. WIREs Comput Stat 11(6):e1483. https://doi.org/10.1002/wics.1483

Zampetakis LA, Moustakis VS (2010) Quantifying uncertainty in ranking problems with composite indicators: a Bayesian approach. J Modell Manag 5(1):63–80. https://doi.org/10.1108/17465661011026176

Zuk O, Ein-Dor L, Domany E (2007) Ranking under uncertainty. In: Parr R, van der Gaag LC (eds) UAI 2007, Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence, Vancouver, BC, Canada, July 19–22, 2007. AUAI Press, pp 466–473

Authors and Affiliations

Balázs R. Sziklai[1,2] · Máté Baranyi[3] · Károly Héberger[4]

✉ Balázs R. Sziklai
sziklai.balazs@krtk.hun-ren.hu

Máté Baranyi
baranyim@math.bme.hu

Károly Héberger
heberger.karoly@ttk.hu

[1] Institute of Economics, HUN-REN Centre for Economic and Regional Studies, Budapest, Hungary

[2] Department of Operations Research and Actuarial Sciences, Corvinus University of Budapest, Budapest, Hungary

[3] Department of Stochastics, Budapest University of Technology and Economics, Budapest, Hungary

[4] Plasma Chemistry Research Group, Institute of Materials and Environmental Chemistry, HUN-REN Research Centre for Natural Sciences, Institute of Excellence, Hungarian Academy of Sciences, Budapest, Hungary